Educ Asse Eval Acc (2011) 23:41–63
DOI 10.1007/s11092-010-9107-x

Teachers matter: Measures of teacher effectiveness in low-income minority schools

Elaine M. Silva Mangiante

Received: 25 May 2010 / Accepted: 1 September 2010 /
Published online: 14 September 2010

Abstract Teachers make a difference in student academic growth. Students from low-income, minority communities attend schools with less resources and less qualified teachers than students in wealthier communities. The Race to the Top (RTTT) policy by the U.S. Department of Education has attempted to address the achievement gap based on SES and the disparity in the quality of teachers between communities. The policy stipulates that teacher effectiveness be determined, in significant part, by student growth measures and supplemented with multiple observation-based assessments. The emphasis placed on student outcomes to indicate teacher effects has served to link teacher evaluations with teacher effectiveness. This review article examines the reported benefits and critical responses to the use of a prominent student growth measure, the Education Value-Added Assessment System (EVAAS), in terms of its implementation as an evaluation tool of teacher effectiveness in low-income, minority schools. Models of observational teacher evaluations, taking into consideration common attributes of effective teachers in low-income schools, are presented as supplemental measures to provide more in-depth information to interpret value-added analyses and to minimize possible misinterpretation of student growth data or the misclassification of teachers' effectiveness for teachers in low-income schools. Information obtained from a combination of evaluation measures can be used to identify both effective and ineffective teachers, to target areas in need of improvement to increase teacher effectiveness, and to make decisions concerning the equitable distribution of effective teachers, especially for students who are most in need.

Keywords Teacher effectiveness · Teacher evaluation · Low-income minority schools · Student-growth measures · Observation-based evaluations

E. M. S. Mangiante (✉)
University of Rhode Island, School of Education, Chafee Hall 705, Kingston, RI 02881, USA
e-mail: emangiante@cox.net

1 Introduction

Teacher quality has become a focus of educational policy in the 21st century. A growing body of research has shown that teacher effectiveness is a strong predictor of differences in student achievement (Ferguson 1998; Sanders and Rivers 1996; Wright et al. 1997; Darling-Hammond 2000). Sanders and Rivers (1996) found that students assigned to ineffective teachers over the course of several years demonstrate significantly lower academic achievement than those students who are assigned to several highly effective teachers in a row indicating that teacher effects on students are both additive and cumulative. Research on teacher allocation policies has documented that schools serving minority, low-income students have the most difficulty recruiting and retaining experienced and effective teachers (Darling-Hammond 2004; Ascher and Fruchter 2001; Krei 1998; Ferguson 1998). African American students are more likely to be assigned to ineffective teachers than white students (Sanders and Rivers 1996; Sanders 2004) and less advantaged schools tend to have more underqualified staff (Darling-Hammond 2000; Lankford et al. 2002). This disparity in teacher quality between schools and districts contributes to a student achievement gap based on SES (Darling-Hammond 2004). Yet, Nye et al. (2004) have reported that teacher effects are larger in low-SES schools in contrast to high-SES schools suggesting that "it matters more *which* teacher a child receives" in a low SES school (p. 254). Nye et al. argue that the range of teacher effectiveness is broader in low-SES schools; thus, efforts to replace less effective teachers with more effective teachers can have a greater effect in low-SES schools than in high-SES schools. In light of these findings, recent educational policies, *No Child Left Behind* and *Race to the Top*, have attempted to address teacher quality (US Department of Education 2002, 2009b, March 7).

The authorization in 2001 of Title 1 of the Elementary and Secondary Act, commonly referred to as No Child Left Behind (NCLB) specified that teachers must be "highly qualified" to ensure that all students learn and demonstrate academic proficiency. Section 1119 of NCLB states that each State must determine measurable objectives "to ensure that all teachers teaching in core academic subjects in each public elementary school and secondary school are highly qualified" (US Department of Education 2002, pp. 1505–1506). To be considered highly qualified, a teacher of a core academic subject is required to hold a bachelor's degree, have full state certification or licensure, and demonstrate subject matter competence (US Department of Education 2004). This focus from NCLB for determining the quality of a teacher is based on teacher preparation—the qualifications or *inputs* from a teacher's training.

This concept of teacher quality was expanded upon at the national level with The American Recovery and Reinvestment Act of 2009: Saving and Creating Jobs and Reforming Education, to focus on teacher effectiveness in terms of *outcomes* of what teachers are able to do to improve student achievement. One of four principles that guide the distribution of funds from this act is to "improve students' achievement through school improvement and reform" (US Department of Education 2009b, March 7, p. 1). This principle, referred to as Race to the Top (RTTT), addresses four specific areas, one of which includes "making improvements in teacher effectiveness and the equitable distribution of qualified teachers for all students, particularly

students who are most in need" (US Department of Education 2009b, March 7, p. 1). As part of this act, an "effective teacher" is defined as:

> a teacher whose students achieve acceptable rates (e.g., at least one grade level in an academic year) of student growth…States, LEAs, or schools must include multiple measures, provided that teacher effectiveness is evaluated, in significant part, by student growth…Supplemental measures may include, for example, multiple observation-based assessments of teacher performance. (US Department of Education 2009a, November, p. 11).

A "highly effective teacher" is indicated through additional measures of evidence: "leadership roles (which may include mentoring or leading professional learning communities) that increase the effectiveness of other teachers in the school or LEA" (US Department of Education 2009a, November, p. 11).

Through a competitive grant program, the RTTT fund has provided an unprecedented $4.35 billion for states that develop ambitious, comprehensive, and achievable plans to improve student outcomes and close the achievement gap (US Department of Education 2010). Of the six requirements for grant applications, the section, *Great Teachers and Leaders*, is assigned the greatest number of possible points. This section addresses the design and implementation of evaluation systems for teachers and principals that "differentiate effectiveness using multiple rating categories that take into account data on student growth…as a significant factor" and use the evaluations to inform human capital decisions such as: professional development, compensation, promotion, retention, tenure, and removal (p. 34). In contrast to NCLB's focus on teacher qualifications, RTTT focuses on teacher effectiveness in terms of teachers' impact on student achievement. The result is a linkage between teacher evaluations and teacher effectiveness.

RTTT calls for teacher effectiveness to be determined from a combination of measures using both students' growth indicators and observation-based assessments. Student growth is based on "the change in student achievement for an individual student between two or more points in time" (US Department of Education 2009a, November, p. 14) rather than on raw student achievement or proficiency data. However, RTTT's inclusion of supplemental measures for teacher evaluation is written in vague language with only one example given: multiple observation-based assessments of teacher performance. Thus, the identification of research-based observation measures to determine teacher effectiveness could be useful for schools to unpack student growth data and to reinforce or disconfirm student growth indicators.

To shed light on the enactment of RTTT measures in schools serving low-income minority populations, this article considers the question of how administrators can blend both student growth measures and observation based assessments to provide an equitable system of teacher evaluation that is workable. To address this question, it is the purpose of this review article (a) to examine reported benefits and critical reviews of one prominent example of a student growth measure, the Education Value-Added Assessment System (EVAAS), as a means to identify effective teachers in low-income, minority schools; (b) to present observation-based approaches used to determine the effectiveness of teachers in schools serving low-income, minority populations; and (c) to present research-based attributes of teachers who have been

found to be effective in promoting student achievement in low-income minority elementary schools. The rationale for examining these three areas is to consider their combined use as a more inclusive means to evaluate and identify effective teachers for the RTTT goal of closing the achievement gap.

2 Value-added models—student growth measures

Currently there has been a research focus on the use of longitudinal data to determine the *value added* to students' growth in achievement by a teacher over the course of a year (Meyer 1997; Sanders 2000; McCaffrey et al. 2004; Raudenbush 2004; Wright et al. 2010). This approach, referred to as "value-added modeling" (VAM), has been developed as an additional measure or as an alternative to reporting student test results as average test scores or percentage of students achieving proficiency on state standards that are required by NCLB to determine schools' Annual Yearly Progress (AYP). Sanders (2003) highlighted the unintended consequences of legislation focused on proficiency scores, particularly for schools serving low-income minority populations. From analyses of several data sets, Sanders found that in response to the federal pressure to increase the percentage of students achieving proficiency, educators, particularly in failing schools, focus their instruction on students who are closest to achieving proficiency while overlooking both high and low achieving students. The result is that the lower achieving students do not progress and the higher achieving students have "suppressed growth" which could impact future academic success (p. 1). Sanders cautioned that an accountability system based on reporting student proficiency percentages will not detect this problematic practice of suppressing the growth of certain student populations. In addition, differences in student outcomes among schools and districts measured by raw mean or proficiency scores are strongly correlated with demographics and socio-economic levels of the students within a school or district (Sanders 2000; Ballou 2002) indicating an achievement gap based on SES (Sanders 2004). Teachers do not have control of the achievement level, socio-economic status, or home environment of their incoming students. Yet, raw test averages are confounded with these factors outside of the control of the teacher and do not indicate either the growth in achievement that a student may make with a particular teacher or the effectiveness of the teacher (Sanders 2000; Raudenbush 2004). For these reasons, Sanders (2003) recommended that states utilize a value-added accountability system to monitor student progress.

One of the most prominent value-added models, the Tennessee Value-Added Assessment System (TVAAS), was originally developed by William L. Sanders and associates at the University of Tennessee (Sanders and Rivers 1996). In 2000, Sanders joined with SAS®, a software company specializing in performance analysis, to build on TVAAS and create the Education Value-Added Assessment System (EVAAS) as a value-added metric to measure schooling influence on student academic success over time (SAS n.d.). VAMs were designed to assess student "*progress* rather than the percentage of students able to meet an absolute standard" (Ballou et al. 2004, pp. 37–38). This measurement of progress is achieved by introducing a prior test score as a student's own starting point to be compared with

post-test scores. If the scales of measure are highly correlated with curricular objectives, reliable, and measure both very low and very high achieving students; the difference in the scores would indicate the contribution by the teacher to student academic growth for that given year (Sanders 2000, 2003; Ballou et al. 2004). By measuring for student gain within a year, "each student serves as his or her own control" (Sanders 2000, p. 333). This control is achieved by comparing test score data for a given student with the student's own prior scores to determine academic growth rather than making comparisons with the performance of other students.

Based on the availability and characteristics of student test data as well as the target of analysis (district, school, or classroom level), different EVAAS models can be implemented (Wright et al. 2010). For the classroom teacher model, when comparably scaled test scores are available for all students in all academic subjects over the course, usually, of 5 years, the multivariate response model (MRM) is used. The MRM model includes test data of students' accumulated learning from previous teachers resulting in a "layered" model (p. 7). The effect on student growth from an individual teacher is interpreted as that teacher's "deviation from the average gain for the district as a whole" (p. 9). However, when the data are not comparably scaled or if pre- and post-data do not exist for a given course, then the univariate response model (URM) can be used to estimate teacher effect for a given year based on a composite of at least three prior test scores.

VAMs, such as the EVAAS, are currently being explored at both the national and state level as a means to determine student growth and, subsequently, the effectiveness of teachers. In this high-stakes accountability climate, researchers and practitioners are posing questions about the role of VAMs in determining teacher effectiveness for evaluative purposes. Do student growth scores account for the complexity of factors influencing teachers and students in the low-income, minority environment? Do student growth scores accurately portray the effectiveness of the teacher and the extent of student learning? The next section will present the benefits gained from using VAM analysis and the arguments against its use as a teacher evaluation tool in low-income, minority schools.

2.1 Considerations when using VAMs in low-income, minority schools

The EVAAS models are well recognized VAMs and regarded as an improvement over mean score assessments; however, several researchers and statisticians have critiqued different aspects of the design (Ladd and Walsh 2002; Ballou 2002; Kupermintz 2003; Amrein-Beardsley 2008; Ishii and Rivkin 2009). To consider both sides of this issue, critiques will be presented in conjunction with responses by EVAAS developers to examine the factors and impact of using VAMs to evaluate teacher effectiveness for students in low-income, minority schools.

2.1.1 Factors affecting student achievement beyond the control of the teacher

There is debate about whether VAMs should include adjustments for SES or demographic variables that are outside the control of the teacher. The developers of EVAAS maintain that use of student growth scores to determine teacher effectiveness eliminates the influence of background factors such as race, ethnicity,

or student ability that bias mean score test data (Ballou et al. 2004). Sanders (2000) asserted that due to the statistical complexity of the EVAAS methodology, student gain scores control for students' backgrounds such that teacher effects obtained from EVAAS models are unrelated to socio-economic factors. One rationale for not including SES or ethnic variables as predictors of student growth was to prevent schools or districts from unwittingly setting inappropriate expectations for students that were not in alignment with the ability of some of its student population. For example, Sanders explained that schools could set low expectations for students from disadvantaged backgrounds that would negatively impact high-achieving students. To address the possibility of misclassification of a teacher's effectiveness based on student growth scores, the layered approach of the MRM was designed to link 5 years of student data with current and previous teachers in order to protect a teacher's evaluation from unexpected changes in students' performance in a given year due to factors such as illness or a family crisis.

However, this argument implies that student gains are a direct result of teacher effectiveness without taking into consideration individual students' motivation level, ability to learn, or access to educational advantages. Berk (1988) posits that the perception of test scores as being objective is "illusory" because the "inferences drawn from [the] scores….can be erroneous, inasmuch as student achievement is not attributable solely to the teacher" (p. 347–348). There are several factors beyond the control of the teacher that affect student achievement. A relevant covariant unaccounted for by VAMs is the motivational levels of the students (Rubin et al. 2004). Does an unmotivated student learn as much as a motivated student in the same class despite the efforts of the teacher? Amrein-Beardsley (2008) argues that students' ability level can affect the amount of yearly growth. When comparing the student growth scores of two classes with equal achievement levels at the beginning of the year but with greatly differing IQ levels, the class with more able students will make greater progress in 1 year; thus, not fairly indicating the effectiveness of the teacher. In evaluating VAMs, Ladd and Walsh (2002) agree that using gains in student achievement is superior to using averages; however, they found that student SES correlates with gains even after prior achievement is accounted for. They caution that policy makers may "attribute the lackluster gain in achievement of a school's students to ineffective teachers…when in fact there may be other explanations such as inadequate resources or other factors outside the immediate control of the school personnel" (Ladd and Walsh 2002, p. 16). Socioeconomic and demographic factors can affect not only the starting point of a student's achievement, "but also their rate of progress" (Ballou 2002, p. 14). At the school level, the unequal availability of resources and well-trained teachers between schools of different socio-economic levels also affects a student's "opportunity to learn" based on the student's address (Marzano 2000). Student factors and availability school resources could result in an inaccurate portrayal of teacher effects on student growth in low-income schools.

In response to these arguments, researchers have empirically confirmed Sanders' (2006) findings that there is a very small to near zero correlation of teacher effects with student-level SES and demographic variables when using EVAAS longitudinal models in comparison with the negative correlations found from the Class Average Score model, Class Average Gain model, or the One-Predictor Fixed Effect

ANCOVA model (Ballou et al. 2004; Lockwood and McCaffrey 2007). The repeated student measures from different subjects and grades "inherent to longitudinal data provide opportunities to control for…unmeasured heterogeneity" such as race/ethnicity and SES (Lockwood and McCaffrey 2007, p. 224). In addition, Sanders and Wright (2008) point out that schools serving poor and minority communities have a higher concentration of novice teachers who are less effective than experienced teachers. They caution that "adjustment for group SES factors will over-adjust the estimates and can camouflage the fact that students in certain schools are not getting an equitable distribution of the teaching talent" (p. 4). This argument suggests that the inclusion of SES variables could hide teacher assignment patterns and teacher effectiveness patterns that may exist in low income, minority schools.

2.1.2 Non-random sorting of students

Researchers have noted that the common school practice of non-random sorting of students can serve as an impediment to estimating a teacher's effectiveness in adding value to a student's academic growth (Berk 1988; Rubin et al. 2004; Ishii and Rivkin 2009). There are several factors that result in the non-random sorting of students in schools or classrooms. The non-random sorting of students between schools is determined by a family's choice of school which is affected by such factors as housing choices, family income, and parental education. Students of low income families who live in high poverty neighborhoods attend schools with less adequate facilities, less qualified teachers, and more regimented curricular instruction than students in wealthier neighborhoods (Kozol 1991, 2005). Darling-Hammond and Post (2000) write "few Americans realize that the U.S. educational system is one of the most unequal in the industrialized world, and students routinely receive dramatically different learning opportunities based on their social status" (p. 127). The data used for school level VAMs do not account for these differences in families or between schools. The varying condition of the quality of schools and their respective teachers based on the wealth of a community can impact the potential for student growth between schools.

Within schools, the common school practice of non-random assignment of students to teachers is a result of principal discretion, parental requests, and teacher choice. Highly effective teachers who are assigned disruptive children or children with greater learning needs are "likely to be evaluated more harshly by the system" (Kupermintz 2003) and be penalized for their circumstances (Ballou 2002; Amrein-Beardsley 2008). Principals may also appease veteran staff by rewarding an experienced teacher with students who are engaged in learning. Conversely, the rating of an ineffective teacher could be inflated if families provide time and resources to support their child's academic growth because they perceive the teacher as inadequate (Ishii and Rivkin 2009). Likewise, the rating of a teacher perceived as being effective could result from families who support their children to become high achievers. Some researchers have suggested that the use of VAMs to evaluate teacher effectiveness combined with the practice of non-random sorting of students could unintentionally produce incentives for teachers to seek employment with higher performing schools or request *high yield* students to avoid serving students from low SES families (Ladd and Walsh 2002; Kupermintz 2003). Ladd and Walsh (2002)

warn that the incentives resulting from this accountability system could "reduce the quality of education in the schools where achievement gains are most needed" (p. 16).

2.1.3 Missing data

Missing student data, a factor that can complicate VAM analyses, is an issue that has received considerable attention. Schools serving student populations in low income, minority communities have high mobility rates (Alexander et al. 1996; Rumberger 2003). Students transferring in from other schools can arrive with incomplete data in their academic files. Another source of missing data results from absences during academic achievement tests. A disproportionate number of low-performing students are absent or are counseled to be absent during standardized tests (Amrein-Beardsley 2008; Rubin et al. 2004). When examining the EVAAS methodology, researchers have questioned whether calculating student growth with missing data may result in giving ineffective teachers an advantage (Kupermintz 2003; Amrein-Beardsley 2008). Ineffective teachers could be given inappropriate credit for the academic growth of low-performing students from incomplete test data used for VAM calculations. Researchers suggest that if VAM measures are used for formal teacher evaluation, then the amount of student data matters in order to provide unbiased estimates of teacher effectiveness (Kupermintz 2003: Raudenbush 2004; Rubin et al. 2004; Amrein-Beardsley 2008).

In response to these critiques, the developers of EVAAS share concern that the non-random pattern of scores that tend to be missing from lower-achieving students can result in selection bias (Sanders and Wright 2008; Sanders et al. 2009). This bias can seriously influence the estimate of the schooling influence on student growth and jeopardize the validity of the analyses. However, to account for this selection bias, Sanders et al. (2009) have explained that the EVAAS multivariate, longitudinal modeling approach incorporates all of the available data in multiple subjects and over multiple years for each student to predict the student's missing scores. In addition, if requirements of data are not met (data are not comparably scaled or either pre- or post-test data are not available for a course), there is a strict requirement that there must be at least three prior scores available for each student in order to conduct VAM analyses using the URM model (Sanders and Wright 2008). However, the usefulness of any statistical analyses is contingent on the quality of the data that are provided.

2.1.4 Validity concerns

VAM statistical analyses depend on the reliability and validity of the test data provided by schools or districts. Several researchers have raised questions concerning the content and construct validity issues with regard to the tests used for student growth measures as an indicator of teacher effectiveness (Andrejko 2004; Reckase 2004; Rubin et al. 2004: Amrein-Beardsley 2008). Does the content of the tests measure what students learn? Are the tests designed to measure the achievement of students well above or well below grade level? Are the pre- and post-scores used to determine student growth measuring the same skills or content?

The EVAAS developers report that they account for these issues by obtaining assurances from schools and districts that the tests are reliable, are highly correlated with curricular objectives, and are able to measure both the very low achieving and very high achieving students (Sanders and Wright 2008).

However, SAS® EVAAS® provides a statistical modeling service and does not have control of test construction. Thus, an important issue to examine before implementing VAM measures is the quality of the assessments. Ravitch (2010) alerts policymakers to this issue and raises questions to consider when choosing to use student growth measures for teacher evaluations.

> If the assessments were low-level, multiple-choice tests, and if teachers were intensely prepping their students for the tests, then could it really be said that these were measures of learning? Or that they were indicators of better teaching? Or were they instead measures of how well children had been drilled to respond to low-level questions? (p. 181).

Ravitch argues that in the current educational climate, use of VAMs to determine teacher effectiveness can result in teachers being evaluated not by curriculum or instruction or "the actual lived experiences of their students," but by data (p. 180). Research has indicated that students living in poverty attending *failing* schools acutely experience instruction for test-taking rather than for learning due to pressure on schools to achieve AYP (Hursh 2007; Darling-Hammond 2007; Ravitch 2010). Based on these findings, it is crucial to consider *what* students are learning and *how* they are learning it to determine if the student growth indicated through statistical analyses is meaningful growth.

Ravitch's argument raises a construct validity issue: Can student growth scores accurately measure teacher effectiveness? Reckase (2004) asserts that the problem lies not in the methods used to determine student growth, but "in the interpretation of the results" (p. 119). For example, for the Effective Practices Incentive Community (EPIC) study of 145 charter schools, Potamites et al. (2009) reported that the Mathematica Policy Research value-added growth distinguished between the top-ranked teacher and the lowest-ranked teacher within the same school; however, teachers in the middle ranking close to each other were not statistically distinguishable. Thus, an important consideration for policymakers is whether the cost of increased testing to provide sufficient data for VAMs is worth the expense when the growth measures identify the most and least effective teachers without distinguishing between teachers in the middle range.

2.1.5 Extent of testing needed

Another factor to consider when using student growth scores as an evaluative tool to determine teacher effectiveness is the extent of testing needed to achieve this goal. VAMs, such as the EVAAS, depend on annually reported longitudinal data for different academic subject areas such as mathematics and reading (Wright et al. 2006). Currently, high stakes testing of selected subjects involves several days to administer and is very costly (Andrejko 2004). Subjects such as music, art, or social studies are not tested, yet there are a substantial number of teachers who teach these unassessed subject areas. The District of Columbia Public School System faced this

challenge in designing the current effectiveness assessment system for school personnel called IMPACT (District of Columbia Public Schools 2009a, b). Two separate assessment protocols were adopted: Group 1 for teachers who teach English and math in grades four through eight for which there are "before" and "after" standardized test scores, and Group 2 for teachers of subjects or grades where value-added data cannot be generated because achievement scores are not available. For Group 1, 40% of a teacher's evaluation is based on observational assessments and 50% is based on the teacher's value-added score; whereas, for Group 2, 80% of a teacher's assessment is based on observation since they do not have value-added scores. This selective testing approach raises the question of "fairness" in using test data to determine the effectiveness of some teachers but not others. It also evokes concerns of whether a requirement to increase the scope of subject testing administered by schools would negatively impact student and teacher morale as well as become a regressive policy adversely affecting schools serving low-income populations with more limited schools budgets than wealthier districts.

2.1.6 Uses of VAMs

At the teacher, school, and district level, VAMs are useful for diagnostic purposes as a measure of progress, a guide for instruction, or as a tool for parents to select schools (Ballou 2002; Andrejko 2004; Raudenbush 2004; Sanders and Wright 2008; Ishii and Rivkin 2009). Sanders and Wright (2008) explain that a multivariate, longitudinal data structure provides a "wealth of positive diagnostic information available for educational decision-makers" (p. 6). Though VAMs are being suggested at the national level for accountability purposes, Sanders and Wright regard the use of VAM data for diagnostic and formative information as having greater importance.

With regard to using VAMs for human capital decisions in schools, some researchers have cautioned that serious difficulties can arise when VAMs are used as a high-stakes evaluation mechanism of teacher effectiveness or as a measure of the effects of instructional practice in schools serving low-income populations (Ballou 2002; Andrejko 2004; Raudenbush 2004; Kupermintz 2003). Schools serving low-income, minority populations face challenges in attracting and retaining quality teachers who provide effective instruction (Darling-Hammond and Post 2000; Darling-Hammond 2004). As a result, one potential problem when attempting to evaluate the effectiveness of teachers using VAMs is the frequent turnover of teachers in schools serving low-income, minority communities. Though a stated benefit of using an EVAAS model is that inequity in the distribution of effective teachers can be detected between advantaged and disadvantaged schools (Sanders and Wright 2008), there will be inadequate data to assess the effects of short-term teachers (Ravitch 2010). In comparison with schools in wealthier districts, schools serving low-income, minority communities face a disproportionate number of challenging factors (i.e. high student mobility, high teacher turnover, missing student data, more test-oriented pedagogy, less access to resources) that could confound the results of value-added models to accurately measure teacher effectiveness. As suggested by Kupermintz (2003), without converging data from teacher, classroom,

and student variables using observational evaluations, misinterpretation of student growth data or the misclassification of teacher effectiveness can result.

3 Call for multiple indicators of teacher effectiveness

To address these problems, researchers have recommended assessing teacher behavior and practice in the classroom through independent evaluations to validate the use of student growth measures to determine teacher effectiveness (Hanushek 1971; Kupermintz 2003; Andrejko 2004; Amrein-Beardsley 2008). To generate evidence of criterion-related validity of VAMs, it is necessary to determine if teachers who post large gains in student growth for a given year are evaluated as highly effective during supervisor evaluations (Amrein-Beardsley 2008). In light of the weight placed on quantitative measures, Andrejko (2004) indicates concern that qualitative assessments of teacher's professional practice will be undervalued despite the standards-based recommendation of using multiple sources of data to assess teacher quality. "What is needed is an active program of research focused on both the development and the evaluation of alternative methods of holding educators accountable" (Koretz 2002, p. 774).

Though VAMs are regarded as an improvement over use of mean scores for assessing student achievement, there are strong voices proposing that value-added scores should not be used in isolation for teacher evaluation without supplemental observational measures (Kupermintz 2003; Andrejko 2004; Amrein-Beardsley 2008). Observational evaluations can yield a deeper understanding of teacher effectiveness *outcomes* based on what a teacher does in the classroom to increase student learning. RTTT calls for supplemental measures to evaluate teacher performance through multiple observation-based assessments in addition to the use of student growth measures. The District of Columbia Public School Department (DCPC) has recently initiated a new teacher effectiveness evaluation system, IMPACT, adopting the dual components of the RTTT recommendations (DCPC 2009a, b). The IMPACT teacher evaluation system incorporates a combination of individual value-added scores, if available, and school value-added scores with ratings from five observations based on teaching and learning standards: three evaluations by a building administrator and two by outside *master educators* who have subject-matter expertise. Through a $10 million Gates grant, Denver is another example of a district that will implement a new system of teacher evaluation using multiple measures including student growth scores and peer observations (Mitchell 2010). To implement this two-pronged approach to teacher evaluation, it is not only important to carefully select a VAM methodology, but also to establish valid observation systems and evaluation criteria of teacher effectiveness in order to identify teachers at both ends of the spectrum—highly skilled, knowledgeable teachers and teachers whose practice is harmful to the academic advancement of students. This section will address the teacher performance criteria as well as the types of evaluators and various observational evaluation systems that have been successfully implemented in different districts to assess teacher effectiveness in low-income, minority schools.

3.1 Teacher performance criteria for observational evaluations

Teachers, as the most important resource in education, need to be knowledgeable in their craft. In order to evaluate the knowledge and skill base of teachers, states and districts have adopted frameworks or developed their own professional teaching standards to assess teacher performance. For example, the research-based *Framework for Teaching* (Danielson 2007), which includes Descriptors of Practice with Levels of Performance, has been adopted by some school districts as valid criteria to develop professional goals and assess teacher effectiveness in the effort to improve student achievement. The framework addresses four main categories of practice with 22 components describing teacher characteristics and behaviors: planning and preparation, the classroom environment, instruction, and professional responsibilities. The development of this framework was informed, in part, by constructivist learning theory and by the National Board for Professional Teaching Standards designed for experienced teachers and it has been correlated with the Interstate New Teacher Assessment and Support Consortium, a framework designed for beginning teachers.

However, some states have developed their own teaching standards. The California Standards for the Teaching Profession (CSTP), designed for both novice and veteran teachers, is organized around six areas of teaching practice (Commission on Teacher Credentialing 2009):

- Engaging and Supporting All Students in Learning
- Creating and Maintaining Effective Environments for Student Learning
- Understanding and Organizing Subject Matter for Student Learning
- Planning Instruction and Designing Learning Experiences for All Students
- Assessing Students for Learning
- Developing as a Professional Educator (p. 3)

Similar to the *Framework for Teaching*, these criteria serve to identify key skills for effective teaching focused on student needs, thoughtfully planned instruction, organization, and professional engagement. The District of Columbia also developed its own Teaching and Learning Framework (TLF) for the IMPACT teacher evaluation program that includes three components: plan (including instruction and the learning environment), teach, and increase effectiveness (DCPC 2009a, b). Since the framework was newly developed, only the *Teach* domain with nine descriptive categories was used for observational assessments of teachers during the 2009–2010 year. This domain focuses on lesson objectives, content delivery, student engagement and understanding, multiple learning styles, positive interaction, and behavior management. In contrast with the Framework for Teaching and the CSTP, there is less emphasis in the TLF on the teacher's professional role in the educational and local community. Instead, the IMPACT program established a separate category for evaluation specifically addressing both core professionalism and commitment to the school community. Clearly defined standards for teacher performance aid evaluators in identifying teacher effectiveness. These reform efforts by some schools and districts have also involved revamping the choice of personnel as evaluators and the structure of evaluation systems.

3.2 Supervisor evaluations of teacher effectiveness

There have been increasing efforts to address the need for multiple measures of teacher effectiveness. Formal teacher evaluations typically have been under the jurisdiction of principals. Empirical literature on subjective evaluations of teacher performance has noted the potential for bias during typical supervisor evaluations from factors such as age and gender of the supervisor and teacher, as well as likeability of the teacher (Wayne and Ferris 1990; Lefkowitz 2000; Varma and Stroh 2001). In an effort to account for potential bias when determining teacher effectiveness, the System for Teaching and Learning Assessment and Review (STAR) process was developed and piloted in urban districts in Louisiana using three assessor types trained in using the STAR: principals, master teachers, and external assessors (Teddlie et al. 1990). The research indicated that the evaluations made by the three assessor types were consistent and correlated. Thus, the assessors had common perspectives as they viewed classroom practice and student learning. However, Teddlie et al. cautioned that since the master teachers gave the highest scores, external assessors were essential to prevent artificially inflated scores because of in-school social relationships between the teachers and the master teacher or principal. In a more recent study, Jacob and Lefgren (2008) found that principals were able to distinguish well between the most and least effective teachers in terms of student achievement gains using subjective evaluations. The principal evaluation ratings were consistent with VAMs for the best and worst teachers. Yet, the principals had more difficulty distinguishing the effectiveness between teachers in the broad middle range. Though this finding indicates that "good teaching is, at least to some extent, observable by those close to the education process" (p. 130), Jacob and Lefgren suggest that policy makers cannot rely solely on principal ratings for fine-grained judgments of teacher performance. They recommend that policy makers incorporate principal assessments of teachers with VAMs to predict future student achievement. Informed by student growth data, principals can observe teacher performance directly in order to address areas in need of improvement such as instruction, curriculum, and classroom management.

3.3 Distributed leadership for teacher evaluations

Recent policy approaches have shifted from the hierarchical authority structure of administrators evaluating teachers to a distributed leadership model where evaluation responsibilities are shared among both administrators and teachers. Research in distributed leadership has indicated positive outcomes for teacher quality and professionalism (Hart 1995). In an effort to guarantee that every child would have an effective teacher, the National Governors Association (NGA) identified goals for teacher evaluation practices reflecting this shift in policy that included incorporating student learning into teacher evaluations, training evaluators, and broadening participation in evaluation design (Goldrick 2002). Teacher evaluations of veteran teachers typically are conducted every 2–4 years without focus on student learning (Goldrick 2002; Hazi and Rucinski 2009). The NGA proposed a purposeful evaluation system that "measures teaching *outcomes*, not simply teaching *behavior"*

(p. 2). The NGA recommended the use of a VAM, if closely aligned with a state's academic standards, in combination with the use of "classroom observation, student work, teacher portfolios, self-evaluation, peer review, and verification of appropriate credentials" to broaden the variety of indicators of a teacher's effectiveness (p. 5).

3.3.1 Multi-faceted standards-based, performance-based evaluation system

One example of a distributed leadership model is the multi-faceted teacher evaluation system which evolved at Vaughn Elementary School, an urban charter school in the Los Angeles Unified School District, as a result of the school's participation in the Teacher Compensation Project with the Consortium for Policy Research in Education (CPRE) at the University of Wisconsin-Madison (Kellor 2005). Beginning in 1999, all teachers were evaluated according to the new evaluation system that was based on the average of the teacher's self-evaluation, a peer evaluation, and an administrator's evaluation for three levels of competencies. For Level 1, teachers were evaluated in three core areas of literacy, ESL, and mathematics four times a year by two evaluators. For Level 2 or 3, teachers were observed in five core areas (literacy, ESL, mathematics, social studies, and science) up to 2 h by two observers, three or four times a year. Level 3 indicated teachers with the greatest competencies and effectiveness. Over the course of 6 years, the evaluation system evolved to include the formation of a Performance Assessment Review (PAR) Committee to oversee the evaluation system; the creation of a pre- and post-conference system, an appeals process, an inter-rater reliability process; and the development of core area performance rubrics for elementary and middle school teachers.

The teacher scores generated from Vaughn's performance-based teacher evaluation system were compared with student growth scores in literacy, mathematics, language arts, and a composite measure from the SAT9 test using a value-added analysis (Gallagher 2004). Results indicated that there was a statistically significant relationship between teacher performance-based evaluation scores and student growth in reading and a positive, but not statistically significant, relationship for mathematics. From interviews and document analyses, Gallagher reported that these differing findings may have resulted from both teachers and evaluators having less knowledge and lower alignment of standards and assessments in mathematics than in reading. Gallagher makes the case that the validity of teacher evaluation systems can be improved significantly by "using subject-specific evaluations conducted by evaluators who have expertise in instruction of the subject they are evaluating" (p. 104). The results of school-wide evaluations using subject specific teacher experts could more accurately inform the professional development needs of individual teachers and of a school faculty as a whole.

3.3.2 Peer assistance and review

Research has indicated inadequacies with typical teacher evaluations by administrators. Principals/administrators lack training in evaluating teachers (Loup et al. 1996) and they rarely give teachers negative evaluations in order to avoid conflict

despite the national focus on teacher quality (Tucker 1997). In 1999, California instituted the Peer Assistance and Review (PAR) program to involve teachers identified for excellence as "consulting teachers" (CT). Two functions of the CTs were to mentor and evaluate "participating teachers" (PTs) which included new teachers as well as veteran teachers with unsatisfactory evaluations from their principals (Goldstein 2003, p. 398). Principals still evaluated the vast majority of the veteran teachers whose evaluations were satisfactory. The CTs were released from teaching duties for 2–3 years to serve as "PAR coaches" to provide both formative support and summative evaluation of the performance of the PTs (Goldstein 2009, p. 895). The CTs were supported and held accountable by a PAR oversight panel chaired by the teacher union president with four teachers and four administrators as panel members. Unlike principals, the coaches were involved in teacher evaluation full time, received training, and engaged in dialogue with other coaches to sort out what the teacher professional standards would look like in practice. For transparency, the CTs were required to present to the panel documentation of a teacher's performance according to the standards for teacher professionalism, documentation of the CT's intervention and support for the PT, and a rational for the future employment recommendations. According to the RTTT definitions, the CTs, in their role as mentors, would be considered "highly effective teachers" (US Department of Education 2009a, November, p. 11).

One key finding from Goldstein's (2009) study of the PAR program in one urban K-12 school district in California was that "10%–12.5% of beginning teachers and almost all veterans across the first 4 years of the PAR program…were removed from classroom teaching" (p. 895) in contrast to the national norm for teacher dismissal of 0.1% (Tucker 1997). Goldstein (2003) describes that both principals and teachers in their first year as a CT were reticent to give negative evaluations. However, the PAR system with the oversight panel had two significant features which supported the panel in making the tough decisions to identify and dismiss ineffective teachers: 1) an established forum for communal dialogue allowed the team to form "a professional community focused on the examination of practice" and 2) the requirement to provide evidence through documentation allowed the conversation to focus "on observed behaviors rather than opaque evaluator opinion" (Goldstein 2009, p. 917). The PAR system depersonalized the process placing both administrators and teachers on the same side of the table focused on teaching practice. According to the teacher union president, "We're trying to institute standards for teaching, so that people will be playing on a common playing field, with common rules. Hiring and firing decisions…would be based on standards rather than the whim of a particular individual" (p. 921). The administrators and coaches, together, were working to improve the quality of teaching in the school. It is noteworthy that dismissal decisions of ineffective new teachers is supported by research that "a teacher who performs poorly in the first 2 years is unlikely to undergo a radical transformation in year three" (Jacob 2004, p. 146). The PAR system provided a means to identify and dismiss those teachers who were ineffective and whose practice was detrimental to the academic advancement of their students.

The teachers as CTs were trusted to conduct teacher evaluations and their recommendations carried weight from their clearly documented evidence grounded

in standards of professional teaching practice. By monitoring their own profession, they were holding their fellow teachers to high standards. The CTs were able to evaluate other teachers and know what dispositions or pedagogical practices were missing for effective teaching. For example, CTs were able to identify teachers who blatantly did not respect students or who were not dedicated to helping students learn. In addition, the CTs could recognize those teachers who consistently struggled with preplanning or organizing lessons despite a year of mentoring and support. Therefore, there were observable attributes of teachers that served as criteria for dismissal of ineffective teachers other than student growth measures.

In summary, there have been lessons learned from successful efforts in various states to reform teacher evaluation systems. The types of personnel responsible for teacher evaluation have evolved from the sole domain of administrators to a distributed model including master teachers and outside experts. Examination of the most successful programs has shown that evaluators are able to assess a teacher's performance more accurately if the evaluator has been trained and if he/she has specialized content knowledge. The newly designed District of Columbia IMPACT program has incorporated this feature into their evaluation system, whereby 31 master educators from a wide range of subject areas and grade levels have been recruited as evaluators (DCPC 2009a, b). In addition, when professional teaching standards are identified, evaluators have clear criteria from which to base their assessments. The California PAR system provided insight into the effectiveness of creating a collaborative team of coaches who provide both formative support and summative evaluation to teachers in coordination with an oversight committee which requires that evaluators document their recommendations with evidence. In terms of the applicability of this type of evaluation program to other urban schools, knowledge of the attributes of teachers who are effective in teaching in low-income, minority schools would further inform PAR coaches in their evaluation process.

4 Characteristics of effective teachers in low-income, minority schools

As the PAR coaches were able to identify both ineffective pedagogical practices and inappropriate dispositions of teachers who were doing harm to their students, researchers have identified common behaviors and observable characteristics of teachers who effectively serve students in low-income, minority schools (Ladson-Billings 1994; Peterson et al. 1991; Lazar 2006; Flynn 2007; Duncan-Andrade 2007; Farr 2010). Data from qualitative studies provide evidence of attributes of effective teachers and their teaching practices. Studies were selected for this discussion based on the following criteria: 1) highly effective teachers in low-income elementary schools were nominated by multiple school personnel and community members, and 2) data were obtained from direct observations of highly effective teachers rather than reported perceptions by educators. Research findings of attributes of effective teachers that really matter for student learning in low-income schools will be discussed including personal beliefs, instructional practices, interpersonal skills, and professional self reflection.

4.1 Beliefs of effective teachers

Findings from case studies of highly effective teachers indicate that two key beliefs held by effective teachers of students in low-income schools recur in the literature: 1) teachers believe their students are competent and capable of excellence and 2) teachers believe it is their role to assume responsibility for their students' achieving excellence (Ladson-Billings 1994; Peterson et al. 1991; Lazar 2006; Duncan-Andrade 2007). Based on a 2-year ethnographic study involving interviews and classroom observations of eight effective teachers selected by community and school personnel nomination, Ladson-Billings (1994) cites the significance of teacher expectations on student achievement. Teachers who have personal views that African-American students have more behavioral problems or are less capable "do not understand that their perceptions of African American students interfere with their ability to be effective teachers for them" (p. 21). She describes that effective teachers not only believe that their students can learn and excel, but also they believe that it is their responsibility to help their students achieve academically. Data collected from multiple documentation sources (i.e. pupil surveys, parent surveys, peer review, pupil achievement data, observation, teacher tests, administration reports) of 12 uncommonly successful urban elementary teachers indicated that effective teachers have high expectations and confidence in their ability to push their students to succeed (Peterson et al. 1991). From a 3-year study of four highly effective teachers in South Los Angeles, Duncan-Andrade (2007) reported that effective teachers can come from various racial, social, or economic backgrounds, but they share a seriousness about their role in aiding all their students to succeed.

4.2 Instructional practices of effective teachers

To achieve this goal of student academic success, effective teachers of low-income, minority students have strong skills in setting clear learning goals, planning and organizing for instruction, delivering well scaffolded lessons, and monitoring students' progress. The academic program is focused and purposefully well defined with readily apparent goals, lesson plans, long term plans, and assessments (Peterson et al. 1991; Farr 2010). These findings are confirmed by research examining the practices of effective elementary teachers in low-income schools who have been successful in helping their students achieve literacy (Lazar 2006; Flynn 2007). Effective teachers implement routines, models, strategies, and carefully constructed approaches to scaffold student learning while also checking for understanding (Lazar 2006; Flynn 2007; Farr 2010). Effective teachers are constantly preparing for their lessons and working relentlessly (Duncan-Andrade 2007; Farr 2010). The intensity of preparation both of pedagogical approaches and subject content knowledge fosters excitement and passion in the instruction they deliver to their students (Ladson-Billings 1994; Duncan-Andrade 2007). An effective teacher is able to create a climate with a clear message that the classroom focus is on the serious work of learning and extending their students' thinking and abilities (Ladson-Billings 1994). Duncan-Andrade reports that this intense

commitment to preparation and instruction results in students being "the top of their schools in traditional measures of student success" even for students who enter the class mid-year, dismissed from other colleague's classrooms (p. 629). This finding indicates that these instructional practices of effective teachers have an effect on student achievement.

4.3 Interpersonal skills of effective teachers

Ladson-Billings (1994) found that expert teachers of African American students not only had more in-depth knowledge of their subject matter, but also more in-depth knowledge of their students than ineffective novice teachers. By teachers' knowing their students well and building relationships with them, the students "developed a greater commitment to learning because of their commitment to their teacher" (p. 125). Effective teachers provide a welcoming environment that creates a sense of trust and community within the classroom (Duncan-Andrade 2007; Farr 2010). Peterson et al. (1991) found that this environment is fostered in part by a teacher's sensitivity, keen observation, and actions to address student needs or problems. It is also fostered by making connections to the students' community through strong communication with parents, bringing students' lives into the classroom, using culturally relevant texts, and inviting the students to talk and write about their culture (Lazar 2006).

4.4 Professional self reflection of effective teachers

Finally, while effective teachers have confidence in their ability to teach, they also engage in frequent self-critique (Duncan-Andrade 2007; Lazar 2006). Reflection allows teachers to assess their own instructional practice as well as their interactions with students in order to identify weak areas and increase their effectiveness. Farr (2010) describes the cycle of reflection as including analysis of outcomes, discernment of causes, and identification/implementation of solutions. Effective teachers in low-income schools use multiple sources of data to purposefully improve their practice. Data sources can include student assessment scores, student work, observations of student engagement, teacher evaluations, parental responses, or videotapes of lessons. To assist them in this reflection process, effective teachers collaborate with colleagues and mentors to share ideas and resources (Farr 2010). Effective teachers are committed to their profession and advancing their skills and abilities.

In summary, inclusion of identified attributes of effective teachers from low-income, minority schools as criteria for teacher evaluation provides evaluators with research-based support to identify both effective and ineffective teachers of students who are most in need of strong instruction. Feedback from observation-based teacher evaluations addressing these key attributes of effective teachers together with the criteria of content and professional teaching standards could assist teachers in low-income, minority schools to understand their strengths and weaknesses. Additionally, clearly defined criteria of effective teaching could also support administrators in making personnel decisions to seek the best possible teachers for students.

5 Conclusion

Research has indicated that the teacher assigned to a student makes a difference in the academic progress made by that student (Sanders and Rivers 1996). Students from low-income, minority communities attend schools with less resources and less qualified teachers than students in wealthier communities (Darling-Hammond 2000, 2004; Ascher and Fruchter 2001; Krei 1998; Ferguson 1998; Lankford et al. 2002; Kozol 1991, 2005). Recent policies by the U.S. Department of Education have attempted to address the achievement gap based on SES and the disparity in the quality of teachers between communities. NCLB was designed to increase teacher quality by stipulating minimum qualifications for teachers, the *inputs* from training that teachers bring to the classroom. However, the focus has shifted with RTTT to improving teacher effectiveness and the equitable distribution of qualified teachers by concentrating on student *outcomes* as a result of individual teacher practice (US Department of Education 2009b, March 7). RTTT stipulates that teacher effectiveness be determined, in significant part, by student growth measures and supplemented with multiple observation-based assessments (US Department of Education 2009a, November). VAMs using longitudinal data, such as the EVAAS, have been developed to determine the *value added* to students' growth in achievement by a teacher over the course of a year (Sanders and Rivers 1996). VAMs are an improvement over mean score assessments to determine student achievement and, subsequently, teacher effectiveness. The methodology employed by EVAAS models is able to adjust for missing student data and provide a value-added score as long as there are at least three prior achievement scores per student. In contrast, calculating the teacher effect for short-term teachers due to the high turnover rate in low-income, minority schools can pose difficulties for schools using VAMs. When evaluating the effectiveness of teachers for students in low-income, minority schools, other factors could confound the value-added scores including the non-random sorting of students assigned to teachers; the student factors of motivation, ability, and access to educational advantages; and the school factor of availability of resources. In addition, the extent of testing required to implement VAMs and the validity of the tests used to measure student learning are other factors to consider when schools choose to use VAMs. Researchers have recognized the usefulness of VAMs for diagnostic purposes as a measure of student progress or as a guide for instruction (Ballou 2002; Andrejko 2004; Raudenbush 2004; Kupermintz 2003); however, they caution against using VAMs as a sole accountability mechanism of teacher effectiveness in schools serving low-income populations due to the confounding factors. An unintended result of using VAMs for high-stakes accountability of teachers could result in teachers seeking transfers from schools serving low SES populations or determining ways to be assigned *high yield* students. The students most in need will become the victims of this system to quantify teacher effectiveness.

As a balance for student growth data, RTTT calls for supplemental observation measures to determine teacher effectiveness. Critical reviews of traditional observation protocols have exposed the lack of training and limitations of administrator evaluations; thus, evaluations from observations may hold a low-profile position in the high-stakes discussion of measures to determine teacher

effectiveness. However, recent evaluation practices, such as the PAR system in California, have adopted a distributed leadership model for evaluating teachers by skilled mentor teachers that holds promise for actively monitoring the effectiveness of teachers and providing a means to dismiss those teachers whose practice is harmful to students (Goldstein 2003, 2009). Observation systems using distributed leadership can provide a qualitative model for teacher evaluation not only based on clearly defined professional teaching standards, but also based on teacher attributes determined to be essential in meeting the needs of students in low-income, minority schools.

This article proposes blending VAMs with observation-based evaluations to provide a fair and informed assessment of teachers in low-income, minority schools. Value-added scores can give teachers an indication of student growth within their classroom and provide incentive to examine curriculum and instruction *if* these scores are used as a guiding tool rather than as a high-stakes gavel. Additionally, evidence from observational evaluations can be used to provide rich information to interpret value-added analyses and serve to minimize the misinterpretation of student growth data or the misclassification of teachers' effectiveness for teachers in low-income schools. Efforts to provide a balanced means to assess teachers using both growth measures and observational evaluations support the goal of RTTT to determine teacher effectiveness in terms of *outcomes* of what teachers are able to do to improve student achievement. Information obtained from both measures can be used to identify effective as well as ineffective teachers, to target areas in need of improvement to increase teacher effectiveness, and to make decisions concerning the equitable distribution of effective teachers, especially for students who are most in need. To advance these efforts, additional research is needed 1) to refine VAMs to address confounding variables prevalent in low-income, minority schools and 2) to create observational evaluation procedures or redesign existing protocols that consider training of evaluators and well-defined criteria to assess in-service teacher effectiveness in low-income, minority schools. In light of the uneven availability of high quality teachers in schools serving low-income populations, it matters not only to identify the effectiveness of the teacher a child receives, but also to create educational policy that truly supports the placement and retention of effective teachers in the neediest of schools in America.

References

Alexander, K. L., Entwisle, D. P., & Dauber, S. L. (1996). Children in motion: school transfers and elementary school performance. *Journal of Educational Research, 90*(1), 3–12.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*(2), 65–75.

Andrejko, L. (2004). Value-added assessment: a view from a practitioner. *Journal of Educational and Behavioral Statistics, 29*(1), 7–9.

Ascher, C., & Fruchter, N. (2001). Teacher quality and student performance in New York City's low-performing schools. *Journal of Education for Students Placed at Risk, 6*(3), 199–214.

Ballou, D. (2002). Sizing up test scores. *Education Next, 2*(2), 10–15.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–65.

Berk, R. A. (1988). Fifty reasons why student gain does not mean teacher effectiveness. *Journal of Personnel Evaluation in Education, 1*, 345–363.

Commission on Teacher Credentialing (2009, October). *California Standards for the Teaching Profession (CSTP)(2009)*. http://www.ctc.ca.gov/educator-prep/standards/CSTP-2009.pdf. Accessed 27 August, 2010.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria: ASCD.

Darling-Hammond, L. (2000). Teacher quality and student achievement: a review of state policy evidence. *Education Policy Analysis Archives, 8*(1), 1–44.

Darling-Hammond, L. (2004). From "separate by equal" to "no child left behind": The collision of new standards and old inequalities. In D. Meier & G. Wood (Eds.), *Many children left behind* (pp. 3–32). Boston: Beacon Press.

Darling-Hammond, L. (2007). Race, inequality and educational accountability: the irony of 'No Child Left Behind'. *Race, Ethnicity, and Education, 10*(3), 245–260.

Darling-Hammond, L., & Post, L. (2000). Inequality in teaching and schooling: Supporting high-quality teaching and leadership in low-income schools. In R. D. Kahlenberg (Ed.), *A notion at risk: Preserving public education as an engine for social mobility* (pp. 127–167). New York: Century Foundation.

District of Columbia Public Schools (2009a). IMPACT The District of Columbia public schools effectiveness assessment system for school-based personnel: Group 1general education teachers with individual value-added data. http://dcps.dc.gov/DCPS/Files/downloads/TEACHING%20&%20LEARNING/IMPACT/DCPS-IMPACT-Group1-Guidebook-September-2009.pdf. Accessed 26 August, 2010.

District of Columbia Public Schools (2009b). IMPACT The District of Columbia public schools effectiveness assessment system for school-based personnel: Group 1general education teachers without individual value-added data. http://dcps.dc.gov/DCPS/Files/downloads/TEACHING%20&%20LEARNING/IMPACT/DCPS-IMPACT-Group2-Guidebook-September-2009.pdf. Accessed 26 August, 2010.

Duncan-Andrade, J. (2007). Gangstas, Wankstas, and Ridas: defining, developing, and supporting effective teachers in urban schools. *International Journal of Qualitative Studies in Education, 20*(6), 617–638.

Farr, S. (2010). *Teaching as leadership: The highly effective teacher's guide to closing the achievement gap*. San Francisco: Jossey-Bass.

Ferguson, R. F. (1998). Can schools narrow the black-white test score gap? In C. Jencks & M. Phillips (Eds.), *The black-white test score gap* (pp. 318–374). Washington: Brookings.

Flynn, N. (2007). What do effective teachers of literacy do? Subject knowledge and pedagogical choices for literacy. *Literacy, 41*(3), 137–146.

Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education, 79*(4), 79–107.

Goldrick, L. (2002). *Improving teacher evaluation to improve teaching quality.* Washington, DC: National Governors Association. http://www.nga.org/cda/files/1202IMPROVINGTEACHEVAL.pdf. Accessed 25 March 2010.

Goldstein, J. (2003). Making sense of distributed leadership: the case of peer assistance and review. *Educational Evaluation and Policy Analysis, 25*(4), 397–421.

Goldstein, J. (2009). Designing transparent teacher evaluation: the role of oversight panels from professional accountability. *Teachers College Record, 111*(4), 893–933.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: estimation using micro data. *The American Economic Review, 61*(2), 280–288.

Hart, A. W. (1995). Reconceiving school leadership: emergent views. *The Elementary School Journal, 96*(1), 9–28.

Hazi, H. M., & Rucinski, D. A. (2009). Teacher evaluation as a policy target for improved student learning: a fifty-state review of statute and regulatory action since NCLB. *Education Policy Analysis Archives, 17*(5), 1–22.

Hursh, D. (2007). Exacerbating inequality: the failed promise of the No Child Left Behind Act. *Race, Ethnicity, and Education, 10*(3), 295–308.

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy, 4*(4), 520–536.

Jacob, B. A. (2004). The challenges of staffing urban schools with effective teachers. *The Future of Children, 17*(1), 129–153.
Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101–136.
Kellor, E. M. (2005). Catching up with the Vaughn Express: six years of standards-based teacher evaluation and performance pay. *Education Policy Analysis Archives, 13*(7), 1–27.
Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources, 37*(4), 752–777.
Kozol, J. (1991). *Savage inequalities*. New York: Crown Publishers.
Kozol, J. (2005). *Shame of the nation*. New York: Three Rivers Press.
Krei, M. S. (1998). Intensifying the barriers: the problem of inequitable teacher allocation in low-income urban schools. *Urban Education, 33*(1), 71–94.
Kupermintz, H. (2003). Teacher effects and teacher effectiveness: a validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 287–298.
Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: getting the incentive right. *Economics of Education Review, 21*, 1–17.
Ladson-Billings, G. (1994). *The dreamkeepers: Successful teachers of African American children*. San Francisco: Jossey-Bass Publishers.
Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: a descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37–62.
Lazar, A. (2006). Literacy teachers making a difference in urban schools: a context-specific look at effective literacy teaching. *Journal of Reading Education, 32*(1), 13–21.
Lefkowitz, J. (2000). The role of interpersonal affective regard in supervisory performance ratings: a literature review and proposed causal model. *Journal of Occupational and Organizational Psychology, 73*(1), 67–85.
Lockwood, J. R., & McCaffrey, D. F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics, 1*, 223–252.
Loup, K. S., Garland, J. S., Ellett, C. D., & Rugutt, J. K. (1996). Ten years later: findings from a replication of a study of teacher evaluation practiced in our 100 largest districts. *Journal of Personnel Evaluation in Education, 10*, 203–226.
Marzano, R. J. (2000). *A new era of school reform: Going where the research takes us*. Aurora: Mid-continent Research for Education and Learning (ERIC document Reproduction Service No. ED454255).
McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, J. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.
Meyer, R. (1997). Value-added indicators of school performance: a primer. *Economics of Education Review, 16*(3), 283–301.
Mitchell, N. (2010, April 9). Building a better teacher in Denver. *Education News Colorado*. http://www.ednewscolorado.org/2010/04/09building-a-better-teacher-in-denver/. Accessed 12 April 2010.
Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257.
Peterson, K. D., Bennet, B., & Sherman, D. F. (1991). Themes of uncommonly successful teachers of at-risk students. *Urban Education, 26*, 176–194.
Potamites, E., Booker, K., Chaplin, D., & Iseberg, E. (2009). *Measuring school and teacher effectiveness in the EPIC Charter School Consortium—Year 2* (Mathematica Reference Number 6325-230), Washington, DC: Mathematica Policy Research, Inc.
Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121–129.
Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.
Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics, 29*(1), 117–120.
Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103–116.
Rumberger, R. (2003). The causes and consequences of student mobility. *Journal of Negro Education, 72*(1), 6–21.
Sanders, W. L. (2000). Value-added assessments from student achievement data: opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*(4), 329–339.

Sanders, W. L. (2003). *Beyond No Child Left Behind.* Paper presented at the 2003 annual meeting of the American Educational Research Association. http://www.sas.com/resources/asset/beyond-no-child-left-behind.pdf. Accessed 22 August, 2010.

Sanders, W. L. (2004). *A summary of conclusions drawn from longitudinal analyses of students achievement data over the past 22 years*. Paper presented at the Governors Education Symposium. http://www.sas.com/resources/asset/hunt_summary.pdf. Accessed 22 August, 2010.

Sanders, W. L. (2006, October 16). *Comparisons among various educational assessment value-added models.* Paper presented at The Power of Two—National Value-Added Conference. http://www.sas.com/resources/asset/vaconferencepaper.pdf. Accessed 26 August, 2010.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Research Progress Report. Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L., & Wright, S. P. (2008). *A response to Amrein-Beardsley (2008) "Methodological concerns about the Education Value-Added Assessment System".* http://www.sas.com/resources/asset/Sanders_Wright_response_to_Amrein-Beardsley_4_14_2008.pdf. Accessed 22 August, 2010.

Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009). *A response to criticisms of SAS® EVAAS®.* http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf. Accessed 22 August, 2010.

SAS (n.d.). *Schooling effectiveness—SAS® EVAAS® for K-12.* http://www.sas.com/govedu/edu/services/effectiveness.html#teachers. Accessed 22 August, 2010.

Teddlie, C., Ellett, C., & Naik, N. (1990, April). *A study of the generalizability of the System forTeaching and Learning Assessment and Review (STAR).* Paper presented at the meeting of the American Educational Research Association, Boston, MA.

Tucker, P. D. (1997). Lake Wobegon: where all teachers are competent (Or, have we come to terms with the problem of incompetent teachers?). *Journal of Personnel Evaluation in Education, 11*, 103–126.

U.S. Department of Education (2002, January 8). *Public Law 107–110, the No Child Left Behind Act of 2001.* http://www.ed.gov/policy/elsec/leg/esea02/index.html. Accessed 4 October 2008.

U.S. Department of Education (2004, March). *New No Child Left Behind flexibility: Highly qualified teachers fact sheet.* http://www2.ed.gov/nclb/methods/teachers/hqtflexibiilty.html. Accessed 25 March 2010.

U.S. Department of Education (2009a, November). *Race to the Top Program Executive Summary.* http://www2.ed.gov/programs/racetothetop/executive-summary.pdf. Accessed 18 February 2010.

U.S. Department of Education (2009b, March 7). *The American Recovery and Reinvestment Act of 2009: Saving and creating jobs and reforming education.* http://www2.ed.gov/print/gen/leg/recovery/implementation.html. Accessed 18 February 2010.

U.S. Department of Education (2010). *Race to the Top application for phase 2 funding* (CFDA Number: 84.395A). http://www2.ed.gov/programs/racetothe top/applicant.html. Accessed 18 August 2010.

Varma, A., & Stroh, L. K. (2001). The impact of same-sex LMX dyads on performance evaluations. *Human Resource Management, 40*(4), 309–320.

Wayne, S. J., & Ferris, G. R. (1990). Influence tactics, affect, and exchange quality in supervisor-subordinate interactions: a laboratory experiment and field study. *The Journal of Applied Psychology, 75*(5), 487–499.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11* (1), 57–67.

Wright, S. P., Sanders, W. L., & Rivers, J. C. (2006). *Measurement of academic growth of individual students toward variable and meaningful academic standards*. Cary, NC: SAS Institute Inc. http://www.sas.com/resources/asset/measurement-of-academic-growth.pdf. Accessed 26 August, 2010.

Wright, S. P., White, J. T, Sanders, W. L., & Rivers, J. C. (2010). *SAS® EVAAS® statistical models*. http://www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf. Accessed 22 August, 2010.